

# Bilinear modulation models for seasonal tables of counts

Brian D. Marx · Paul H.C. Eilers · Jutta Gampe ·  
Roland Rau

Received: 11 March 2009 / Accepted: 11 July 2009 / Published online: 24 July 2009  
© The Author(s) 2009. This article is published with open access at Springerlink.com

**Abstract** We propose generalized linear models for time or age-time tables of seasonal counts, with the goal of better understanding seasonal patterns in the data. The linear predictor contains a smooth component for the trend and the product of a smooth component (the modulation) and a periodic time series of arbitrary shape (the carrier wave). To model rates, a population offset is added. Two-dimensional trends and modulation are estimated using a tensor product B-spline basis of moderate dimension. Further smoothness is ensured using difference penalties on the rows and columns of the tensor product coefficients. The optimal penalty tuning parameters are chosen based on minimization of a quasi-information criterion. Computationally efficient estimation is achieved using array regression techniques, avoiding excessively large matrices. The model is applied to female death rate in the US due to cerebrovascular diseases and respiratory diseases.

**Keywords** Array regression · Lifetable · P-splines · Regularization · Seasonality · Tensor product

## 1 Introduction

Judging from observed monthly counts of deaths (by year and age), many diseases show seasonal patterns. It is of interest to model these patterns, their overall strength and the relative strengths in different months. Our primary interest to better understand the seasonal patterns of deaths due to various diseases and how these patterns vary with year and age. In our earlier work (Eilers et al. 2008), we proposed a modulation model: the pattern within years is described by a (co)sine and its amplitude is described by smooth varying-coefficient surfaces, over year and age. To this a trend is added.

Such a modulation model can describe seasonal patterns quite well, but from studying the residuals it became clear that the (co)sine function is too simple: it cannot handle relatively sharp peaks in winter and relatively flat troughs in summer. One possible solution is to introduce modulated harmonics of double or triple frequency, in the spirit of Fourier analysis. Instead we opt for a different approach, one that offers better opportunities for generalization. Specifically, in a one-dimensional series of counts (of death rates), we assume that there exists a “carrier wave” (a term we borrow from radio technology), which is modulated over time. The period of carrier wave is 12, and it is parameterized by the vector  $\gamma$ . This leads to a bilinear modulation model. We propose an iterative algorithm, cycling between estimation of the carrier wave and the modulation. With proper normalization of  $\gamma$  this leads to an identifiable model that can easily be estimated by iterative weighted least squares.

---

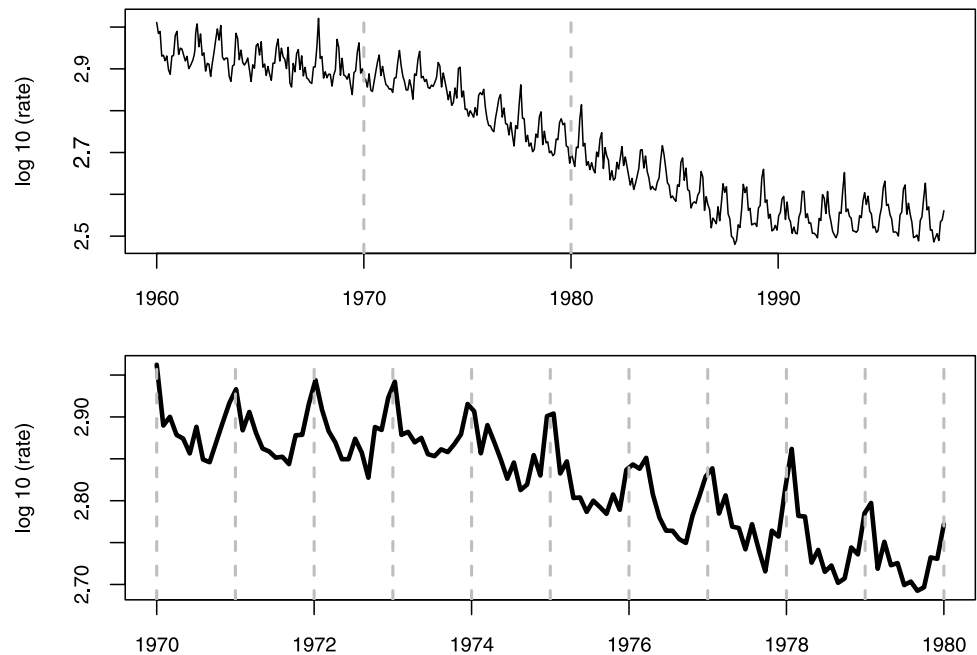
B.D. Marx (✉)  
Department of Experimental Statistics, Louisiana State  
University, Baton Rouge, LA 70803, USA  
e-mail: [bmarx@lsu.edu](mailto:bmarx@lsu.edu)

P.H.C. Eilers  
Department of Biostatistics, Erasmus Medical Center, 3015 GE  
Rotterdam, The Netherlands  
e-mail: [p.eilers@erasmusmc.nl](mailto:p.eilers@erasmusmc.nl)

J. Gampe  
Max Planck Institute for Demographic Research, 10857 Rostock,  
Germany  
e-mail: [gampe@demogr.mpg.de](mailto:gampe@demogr.mpg.de)

R. Rau  
Institute of Sociology and Demography, University of Rostock,  
Rostock, Germany  
e-mail: [roland.rau@uni-rostock.de](mailto:roland.rau@uni-rostock.de)

**Fig. 1** Time series plot of death rate per 100 000 per year due to cerebrovascular diseases for U.S. females during 1960–1998, summed over ages 51–100. The *bottom panel* focuses on 1970–1980



Section 2 first introduces the one-dimensional modulation model for Poisson counts, over time. We start simply with a varying sine/cosine seasonal trend. We find this model to be too simplistic, and develop the bilinear model, which allows for a more general varying seasonal “carrier wave.” A combined model is also presented that allows for a sine/cosine trend in combination with a general carrier wave. Sections 2.3 and 2.4 present a suggestion for optimal choice of the tuning parameters, as well as one-dimensional examples. Section 3 extends the one-dimensional bilinear model into two dimensional, now over both year and age. The Poisson counts are on a regular grid. All of our models are essentially P-spline varying-coefficient models, where in the case of the two-dimensional model tensor product B-splines and an anisotropic difference penalty are used. Our model is computationally demanding. We analyze a table with 50 rows and 468 columns, fitting models with either 169 or 507 tensor product parameters. The standard approach, using vectorization of the data table and Kronecker products for the tensor product basis, puts heavy demands on memory and computation time. Fortunately we can adapt the array regression algorithms (Currie et al. 2006; Eilers et al. 2006) to our case, gaining at least an order of magnitude in efficiency. Details are presented in Sect. 3.1. A two-dimensional application is given in Sect. 4. We close with a Discussion.

## 2 One-dimensional model details

To start simply, we describe the models in one dimension (over time, summing over age). We move into two dimen-

sions in Sects. 3 and 4. Consider Fig. 1 which plots the  $\log_{10}$  monthly death rates (counts/exposures) due to cerebrovascular diseases for females in the United States during 1960 through 1998. The death and exposure counts are summed over the ages 51–100. Death counts are obtained from NCHS’s Multiple Cause of Death Data which can be downloaded from the website of the National Bureau of Economic Research ([www.nber.org](http://www.nber.org)). See our references for more detail. Population counts for single ages for the 1st of January for each year were downloaded from the Human Mortality Database ([www.mortality.org](http://www.mortality.org)). Monthly population figures were estimated via linear interpolation between the years for single ages. The observed data are a series of pairs of counts and exposures,  $(y_t, e_t)$ ,  $t = 1, 2, \dots, T$ , where  $T = 468 = 12 \times 39$  for Fig. 1. The lower panel of Fig. 1 focuses on ten years during the 1970s, providing clearer evidence of strongly varying seasonal patterns, but generally with peaks during winter and troughs during summer.

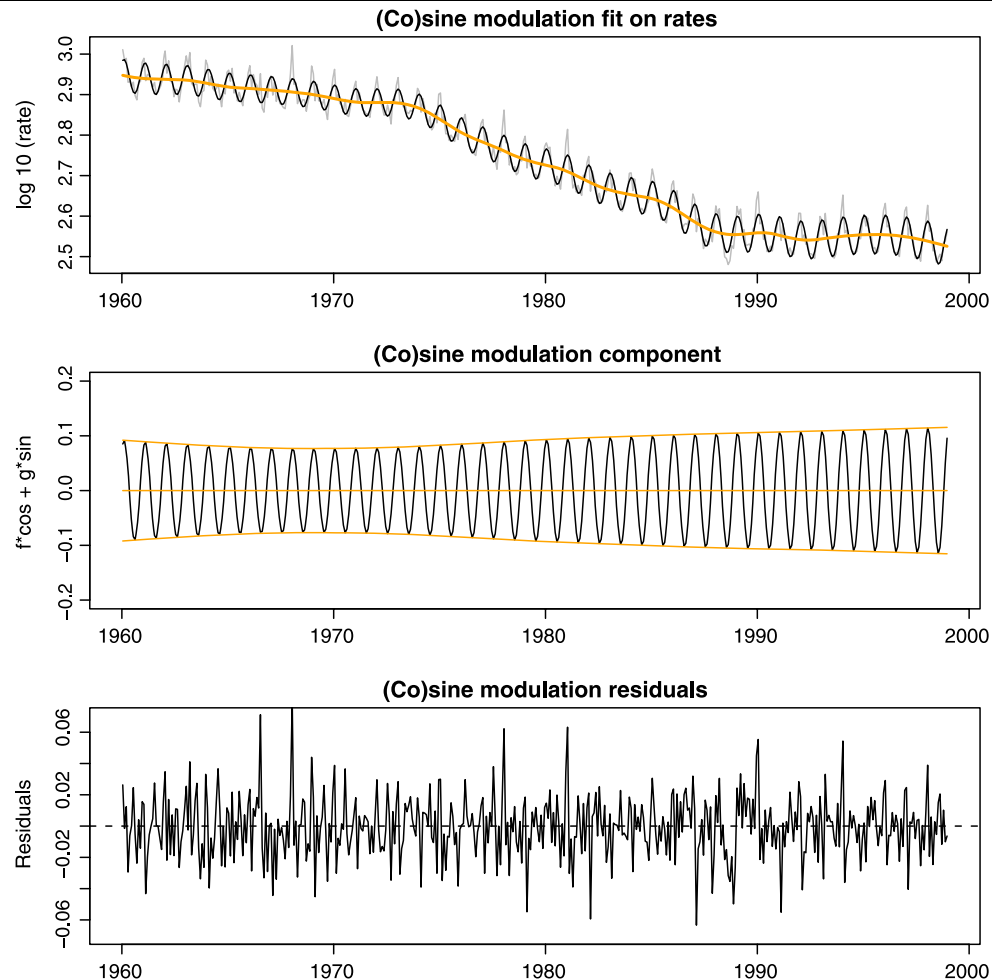
### 2.1 The cosine-sine modulation model

We consider models for the logarithm of an expected value of a Poisson random variable ( $y_t$ ), i.e.  $\log(\mu_t) = \eta_t$ , where  $\mu_t = E(y_t)$ . Eilers et al. (2008) described the linear predictor model

$$\eta_t = \log(e_t) + v_t + f_t \cos(\omega t) + g_t \sin(\omega t), \quad (1)$$

where  $\omega = 2\pi/12$  and the series of varying-coefficients  $v$ ,  $f$ , and  $g$  are smooth, and are modeled using P-splines. For our application, exposures,  $e_t$ , correspond to population

**Fig. 2** P-spline (co)sine modulation model fit and trend for female cerebrovascular death rate (per 100 000 per year) (*top*); the corresponding detrended (co)sine seasonal component (*middle*); the plot of residuals (*bottom*)



size. This is a modulation model: the amplitude of the seasonal waveforms is varied in strength by  $f$  and  $g$ . Technically this is a varying-coefficient model (VCM); the  $3T$  parameters (with  $T$  observations) are estimated using P-splines (Eilers and Marx 2002), requiring three penalty terms in the penalized objective.

We save most of the specifics of P-spline fitting for the two-dimensional extension in Sect. 3. However, estimation of the stacked B-spline varying-coefficients,  $\theta$ , (for trend, sine and cosine regressors) chooses  $\theta$  to maximize the penalized Poisson log-likelihood

$$l^*(\theta) = l(\theta) - \frac{1}{2}P,$$

which yields the iterative “normal” equations

$$(M' \tilde{W}_c M + P) \tilde{\theta}_{c+1} = M' \tilde{W}_c \tilde{z}_c, \quad (2)$$

where the effective regressors are  $M = (B \mid \text{diag}\{\cos(\omega t)\}B \mid \text{diag}\{\sin(\omega t)\}B)$ ,  $B$  is the  $(T \times K)$  B-spline basis (using rich equally-spaced knots),  $z$  and  $W$  are the Poisson “working” response and diagonal weight matrix, respectively;  $c$  represents the current iterate. The vector  $\theta$  consists

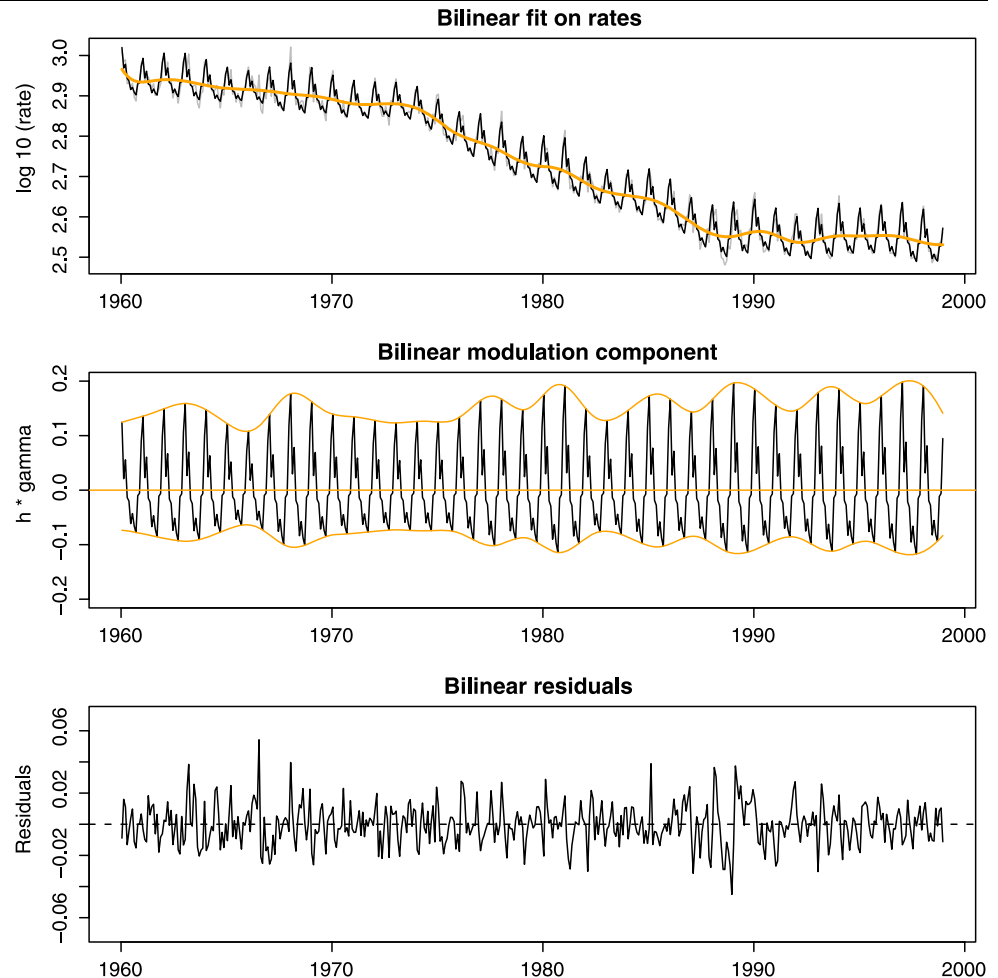
of stacked B-spline coefficients for the varying-coefficients of the intercept (trend), cosine, and sine components, respectively. The intercept (trend) and the varying slopes for the cosine and sine regressors each have their own tuning parameter that are reflected in the block diagonal penalty matrix

$$P = \text{block diag}(\lambda_1 D' D, \lambda_2 D' D, \lambda_3 D' D). \quad (3)$$

The matrix  $D$  is the difference matrix of order  $d$ , and in practice we often set  $\lambda_2 = \lambda_3$ .

Figure 2 (top) overlays the female cerebrovascular death rate data (per 100 000 per year) on a  $\log_{10}$  scale with the (co)sine modulation fit and the trend. This figure also demonstrates the effect of the P-spline tuning parameters on the fit: we have light smoothing on trend and heavy smoothing on the seasonal component. The latter is perhaps more clearly seen in the middle panel, where the seasonal component is detrended. These panels reflect optimal fits; optimization of the tuning parameters will be discussed in Sect. 2.4. The time series of residuals are plotted in the bottom panel  $\log_{10}(y/\hat{\mu}) = \log_{10}(y/e) - \log_{10}(\hat{\mu}/e)$ . The model manifests features which generally are not characteristic of (co)sine functions, as confirmed by regularities

**Fig. 3** P-spline bilinear model fit and trend for female cerebrovascular death rate (per 100 000 per year) (*top*); the corresponding detrended bilinear component (*middle*); the plot of residuals (*bottom*)



among the residuals. Thus, rather than extending the modulation model with sine and cosine waves of double and triple frequency (each having their own modulation), we next motivate the more general bilinear model to capture this residual pattern.

## 2.2 A bilinear model

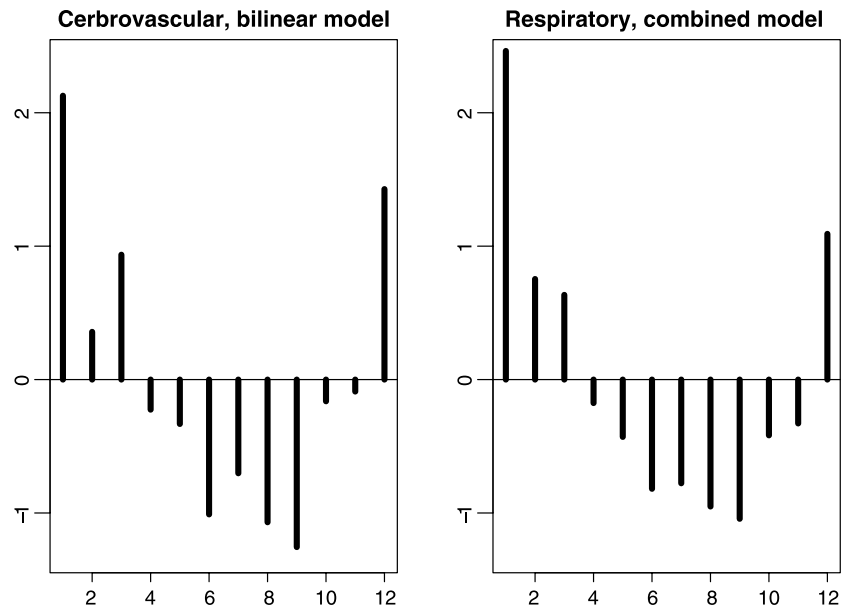
Imagine a more general and unrestricted carrier wave, characterized by the 12-vector  $\gamma$  (one entry for each month). The model is

$$\eta_t = \log(e_t) + v_t + h_t \gamma_{[t]}, \quad (4)$$

where  $h_t$  is the varying-coefficient for the carrier wave. We use the notation  $[t]$  for  $\{(t-1) \bmod 12\} + 1$ , which, simply put, produces an index of 1 through 12 replicated  $T/12$  times. Figure 3 (top) overlays the female cerebrovascular death rate data (per 100 000 per year) on a  $\log_{10}$  scale with the bilinear optimal fit and the trend. The middle panel shows the detrended bilinear component. The time series of residuals are plotted in the bottom panel, again  $\log_{10}(y/\hat{\mu})$ . The residuals generally show a better fit. Consequently there

is more than 35% reduction in the residual standard error, from 29.35 to 18.97, and we find some benefit from generalizing the carrier wave in the bilinear model. Figure 4 (left) shows the shape of the estimated carrier wave for female cerebrovascular disease. Notice the positive spikes in winter months of the carrier wave, which perhaps are tracking epidemics. To a lesser extent, the carrier wave estimates negative spike in the summer months. The local trough in February for cerebrovascular diseases does not seem to be an artifact. A similar dip has been found for Denmark for mortality from all-causes Rau and Doblhammer (2003). It could be the outcome of some “harvesting effect” in January: some of those would have “normally” died in February of cerebrovascular disease already died in January. Alternatively, we can hypothesize that people dying in January died of immediate causes of the cold climate (as shown in the right panel of Fig. 4 for respiratory diseases). Circulatory diseases such as cerebrovascular disease are, however, not only affected by the sudden impact but probably by the accumulation of detrimental effects throughout the entire cold period.

**Fig. 4** Estimated carrier wave for female cerebrovascular bilinear death rate model (*left*), and for female respiratory death rates combined model (*right*)



Unlike the (co)sine modulation approach, the wave is unknown, and hence the model is bilinear because both  $h$  and  $\gamma$  are unknown. We see that  $h_t \gamma_{[t]}$  is a product of functions, and fitting such products has been considered, e.g., by Tutz and Binder (2004). We propose the following iterative algorithm: (i) For given  $\gamma$ , we have a VCM again to estimate  $h$  and  $v$ . Specifically we use the system of equations in (2), now with  $M = (B \mid \Gamma B)$ , and  $P$  having only the first two entries of its block diagonal structure. The (diagonal) matrix  $\Gamma = \text{diag}(c) \otimes \gamma$ , where  $c$  is a vector of ones of length  $T/12$ . (ii) For given  $h$  and  $v$ , estimation of  $\gamma$  can be updated using (generalized) linear regression, which can be viewed as an “estimated” wave basis. We cycle back and forth until convergence: (i) to improve both the  $v$  trend the  $h$  modulation (given  $\gamma$ ), use  $\Gamma B \gamma$ , and (ii) to improve  $\gamma$  carrier wave (given trend and modulation), use  $V R \gamma$ , where  $R = c \otimes I_{12}$  and  $V = \text{diag}(v)$ . Starting values for  $\gamma$  are simple monthly means. Standard identifiability conditions are required on  $h$  or  $\gamma$ ; we choose  $\sum_{k=1}^{12} \gamma_k = 0$  and  $\sum_{k=1}^{12} \gamma_k^2 = 12$ .

### 2.3 The combined model

Despite the generality of the carrier wave, one potentially limiting feature of the bilinear model is that the carrier wave exhibits peaks in the same months from one year to the next. On the other hand, while the (co)sine modulation model has limited flexibility in carrier structure, its peak is allowed to shift. It is natural to try to combine (1) and (2) into one model: the (co)sine then models the major periodic seasonal patterns, while  $\gamma$  and  $h$  would explain additional shocks, i.e.

$$\begin{aligned} \log(\mu_t) &= \log(e_t) + v_t + f_t \cos(\omega t) \\ &\quad + g_t \sin(\omega t) + h_t \gamma_{[t]}. \end{aligned} \quad (5)$$

Implementation of the combined model can be achieved in a similar fashion to the bilinear model, this time switching back and forth between  $\gamma$  and  $(v, s, g, h)$ . Section 2.5 presents the results of the combined model for female respiratory death rates.

### 2.4 Further summary: female cerebrovascular death rates

The co(sine) modulation model, the bilinear model, and the combined model were all fit to the female cerebrovascular death count data (1960–1998, summed over ages 51–100). In all models, the smooth varying coefficient terms used cubic B-splines on  $(25 + 3)$  equally-spaced knots and a second order difference penalty on each set of B-spline coefficients. Since we used a common tuning parameter for both the cosine and sine smooth terms, two tuning parameters are needed for each of the modulation and bilinear models, and three are needed for the combined model.

Table 1 presents the results of the optimal fits based on minimizing a quasi-AIC criterion (Lee et al. 2006; Eilers et al. 2008),

$$\text{QIC} = Q(y, e; \mu) + 2\text{ED} = T + \text{ED} + T \log(\hat{\phi}).$$

The extended (log-)quasi-likelihood is chosen because of evidence of over-dispersion and is defined as

$$Q(y, e; \mu) = \text{dev}(y; \mu) / \phi + T \log \phi,$$

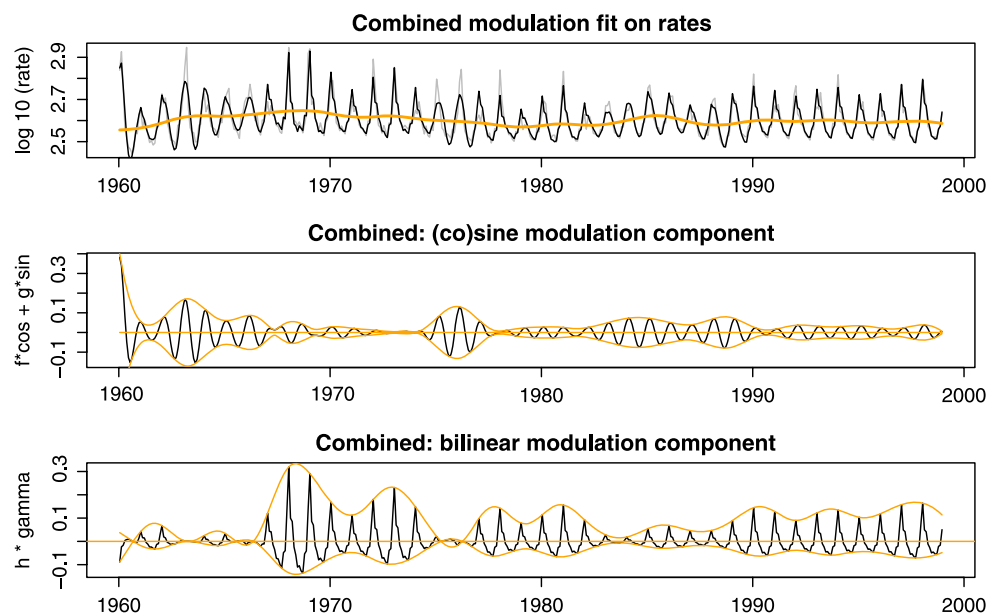
where  $\phi = \text{dev}(y; \mu) / (T - \text{ED})$  and  $T$  is the number of observations. The effective dimension of the model (ED) is approximated by the trace of the smoother matrix. To search for the optimal tuning parameters,  $\log_{10}(\lambda)$  (for each term) was varied on a linear grid, in steps of 0.5. We find that the

**Table 1** Summary: one-dimensional female cerebrovascular models (25 + 3 knots,  $d = 2$  penalty)

Model	QIC	ED	Deviance	Original dimension	Optimal $\lambda$	SE (rate residues)
Cosine-sine	<b>1861</b>	32	7993	1404 (84)	$[10^3, 10^6]$	29.35
Bilinear	<b>1499</b>	51	3405	948 (68)	$[10^2, 10^{2.5}]$	18.97
Combined	<b>1463</b>	75	2806	1884 (124)	$[10^{2.5}, 10^{2.5}, 10^5]$	17.20

**Table 2** Summary: one-dimensional female respiratory models (25 + 3 knots,  $d = 2$  penalty)

Model	QIC	ED	Deviance	Original dimension	Optimal $\lambda$	SE (rate residues)
Cosine-sine	<b>2560</b>	43	33883	1404 (84)	$[10^{0.5}, 10^5]$	53.24
Bilinear	<b>2405</b>	52	23381	948 (68)	$[10^{0.5}, 10^{1.5}]$	45.97
Combined	<b>2371</b>	94	17790	1884 (124)	$[10^3, 10^1, 10^2]$	38.53

**Fig. 5** Combined model for respiratory death rate: model fit and trend (top); sinusoidal component (middle); bilinear carrier component (bottom)

combined model, when compared to the modulation and bilinear models, performs best based on QIC. We also find further evidence that the bilinear model and the combined model are actually quite strong competitors, which is confirmed with the reported standard errors of residuals in Table 1. For cerebrovascular disease, it appears that despite the 35% reduction in residual standard error going from the (co)sine modulation model to the bilinear model, there is not as much additional gain moving to the combined model (<10%). In the next section, we will however see significant benefits in using the combined model for the female respiratory death rate data. Table 1 also shows that ED is naturally higher for the combined model compared to its simpler versions. Conversely, the deviance decreases as we move from the modulation, to the bilinear, and then to the combination model. The original (unprojected) dimension of the coefficient space is also reported, as well as the dimension projected onto B-spline (plus the dimension of the

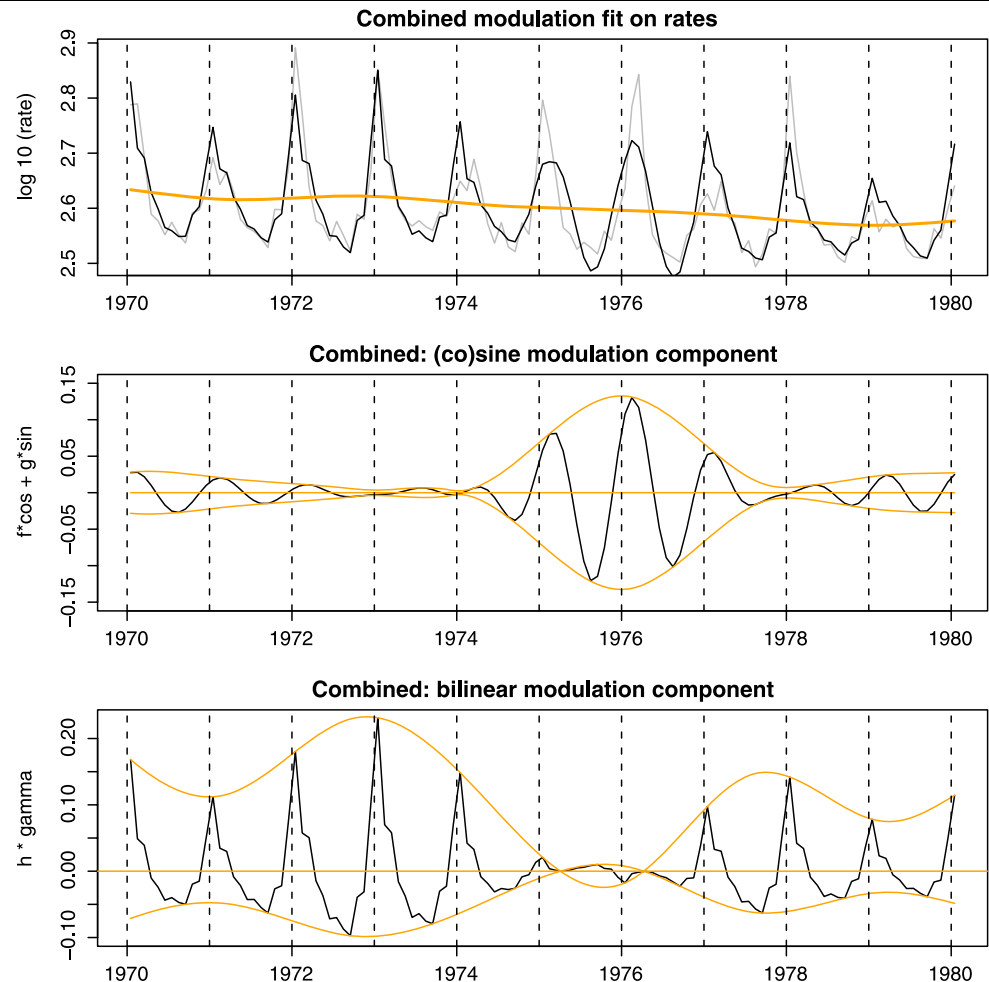
carrier wave). The optimal tuning parameters are provided for reproducibility.

## 2.5 A second example using female respiratory death rate data

The co(sine) modulation model, the bilinear model, and the combined model were also fit to the female respiratory count data (1960–1998, summed over ages 51–100). We find that evidence in favor of the combined model, when compared to the (co)sine modulation and bilinear models. Table 2 provides the summaries for the three models fit to the respiratory data. The last column, corresponding to the standard deviation of the residuals, shows a 13.6% reduction in residual standard error moving from the (co)sine modulation to the bilinear model, and an additional 16.2% reduction moving to the combined model. Figure 5 displays the combined model fit and trend (top) followed by the (co)sine modulation and bilinear components (middle and bottom). Notice



**Fig. 6** Combined model for respiratory death rate: enlarged fit during 1970–1980



how the sinusoidal and bilinear components take turns regularly: when the (co)sine term is high, the bilinear term is not, and vice versa. As noted, the modulated carrier wave must have its peak in the same month (or months). However the (co)sine term can change its phase and so it has a flexible position, i.e. when the peak in a year is not at the carrier peak, it gets a small amplitude and the (co)sine dominates. Figure 4 (right) displays the carrier wave for female respiratory combined model, which has similar features to the carrier wave for female cerebrovascular bilinear model.

Figure 6 enlarges the combined fit during 1970–1980. These are the years where the (co)sine and bilinear terms appear to take turns regularly. The asymmetric carrier wave appears to do its work effectively, with the exception of the years 1975–1977. During the three years, the positions of the peaks shift and the (co)sine tries to correct, by adjusting its phase, for the lack of fit of the bilinear term.

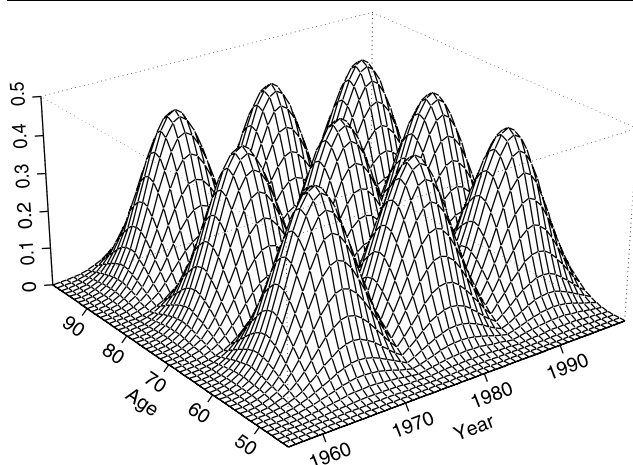
### 3 The two-dimensional bilinear model

Eilers et al. (2008) also formulated the modulation model in two dimensions; here we extend the bilinear model as

such. The counts and exposures are now arranged in a two-dimensional table  $Y = [y_{ta}]$ ,  $E = [e_{ta}]$  e.g., indexed by both time,  $t = 1, \dots, T$ , and by age,  $a = 1, \dots, A$ . The two-dimensional bilinear model is expressed as,

$$\log(\mu_{ta}) = \eta_{ta} = \log(e_{ta}) + v_{ta} + h_{ta}\gamma_{[t]}. \quad (6)$$

Notice that (6) has a double (time and age) subscript for the exposure offset, varying trend, and modulation coefficients, producing varying coefficient surfaces. The carrier has the same periodic structure as with the one-dimensional bilinear model. The varying coefficients  $v$  and  $h$  are assumed to be smooth along time and age, and thus we model them using tensor product B-spline bases, allowing general surfaces. In the spirit of a P-spline approach, we avoid knot selection by: (i) using a sufficiently rich  $K \times L$  gridded tensor product basis, and (ii) imposing penalties on the coefficients associated with the rows and columns of the tensor product basis, where each penalty is regularized by its own positive tuning parameter,  $\lambda$ , hence allowing for anisotropic smoothing. Despite the fact that our penalization allows for anisotropy, we point out that it is only a limited type: one that only allows for anisotropic smoothness along the age and time axes.



**Fig. 7** A sparse representation of a cubic tensor product basis

Figure 7 shows a sparse representation of a cubic B-spline tensor product basis; a figure containing the full basis is too crowded to appreciate its structure. Associated with each mountain in Fig. 7 is an unknown tensor product coefficient, which together can be placed into a  $K \times L$  matrix. Multiplying each mountain by its coefficient and summing the basis produces very general surfaces. Let  $B = [b_{tk}]$  ( $\check{B} = [\check{b}_{al}]$ ) be the  $T \times K$  ( $A \times L$ ) B-spline basis on the time (age) domain. Denote  $\mathcal{A}$  and  $\mathcal{B}$  as the  $K \times L$  matrices of the tensor product coefficients for  $V = [v_{ta}]$  and  $H = [h_{ta}]$ , respectively. We can rewrite (6) as

$$\begin{aligned} \log(M) &= \log(E) + V + \Gamma H \\ &= \log(E) + B\mathcal{A}\check{B}' + \Gamma B\check{B}\check{B}', \end{aligned} \quad (7)$$

where  $M = [\mu_{ta}]$  and  $\Gamma = C \otimes \gamma$ . Again, for fixed diagonal matrix  $\Gamma$ , we have a two-dimensional VCM; for fixed  $V$  and  $H$ , we have a generalized linear regression. Identifiability conditions are also again chosen to be:  $\sum_{k=1}^{12} \gamma_k = 0$  and  $\sum_{k=1}^{12} \gamma_k^2 = 12$ .

Penalties are now applied to both rows and columns of  $\mathcal{A}$  and  $\mathcal{B}$ . Denote the (second order) difference penalty matrices  $D$  and  $\check{D}$  with dimensions  $(K-2) \times K$  and  $(L-2) \times L$ , respectively. Figure 8 provides a visualization of strong row and column penalization, producing linear trends for the tensor product coefficients within the rows and columns. The penalty is defined as

$$\begin{aligned} P &= P_{\mathcal{A}} + P_{\mathcal{B}} \\ &= \{\lambda_1 \|D\mathcal{A}\|_F + \check{\lambda}_1 \|\mathcal{A}\check{D}'\|_F\} \\ &\quad + \{\lambda_2 \|D\mathcal{B}\|_F + \check{\lambda}_2 \|\mathcal{B}\check{D}'\|_F\}, \end{aligned} \quad (8)$$

where  $\|\cdot\|_F$  indicates the Frobenius norm, or the sum of the squares of all elements. The penalty is composed of two parts, e.g. the first is equivalently

$$P_{\mathcal{A}} = \text{vec}(\mathcal{A})'[\lambda_1(I_L \otimes D'D) + \check{\lambda}_1(\check{D}'\check{D} \otimes I_K)]\text{vec}(\mathcal{A}),$$

where  $I$  is the identity matrix. The tensor product coefficients,  $\mathcal{A}$  and  $\mathcal{B}$ , are found by maximizing the penalized Poisson log-likelihood function

$$l^*(\mathcal{A}, \mathcal{B}) = l(\mathcal{A}, \mathcal{B}) - \frac{1}{2}P, \quad (9)$$

where the  $\frac{1}{2}$  is a small trick to absorb the 2 under differentiation.

### 3.1 Efficient computation using array regression

We present the array algorithm in a nutshell here; for a more detailed account one should consult Currie et al. (2006) or Eilers et al. (2006). Assume that the  $T \times A$  matrix  $Z = B\mathcal{A}\check{B}'$  is a model for the expected values of the matrix  $Y$  of the same size. The bases  $B$  and  $\check{B}$  have sizes  $T \times K$  and  $A \times L$  respectively, and  $\mathcal{A}$  is the  $K \times L$  coefficient matrix. A  $T \times A$  matrix  $W$  is given, and to estimate the coefficients by weighted least squares we minimize  $\sum_t \sum_a w_{ta}(y_{ta} - z_{ta})^2$ .

The classical solution (1) vectorizes the matrices:  $y = \text{vec}(Y)$ ,  $\alpha = \text{vec}(\mathcal{A})$  and  $z = \text{vec}(Z)$ ,  $w = \text{vec}(W)$ ; (2) constructs a diagonal matrix  $W^* = \text{diag}(w)$ ; (3) combines the bases with a Kronecker product,  $B^* = \check{B} \otimes B$ ; and (4) minimizes  $(y - B^*\alpha)'W^*(y - B^*\alpha)$ , which is standard weighted linear regression. The normal equations are

$$(\check{B} \otimes B)'W^*(\check{B} \otimes B)\hat{\alpha} = Q\hat{\alpha} = (\check{B} \otimes B)'W^*y. \quad (10)$$

It is easy to see that  $B^*$ , with dimensions  $AT \times KL$ , can require a lot of memory space. Also, but perhaps less obvious, the multiplications and sums that lead to the elements of  $Q$  are rather fine-grained and waste an enormous amount of processing time. Both problems are eliminated by rearranging the computations.

Let  $R = B \square B$  indicate the row-wise tensor product of  $B$  with itself. Hence  $R$  has  $T$  rows and  $K^2$  columns and each row of  $R$  is the tensor product of the corresponding row of  $B$  with itself. One can show that the elements of

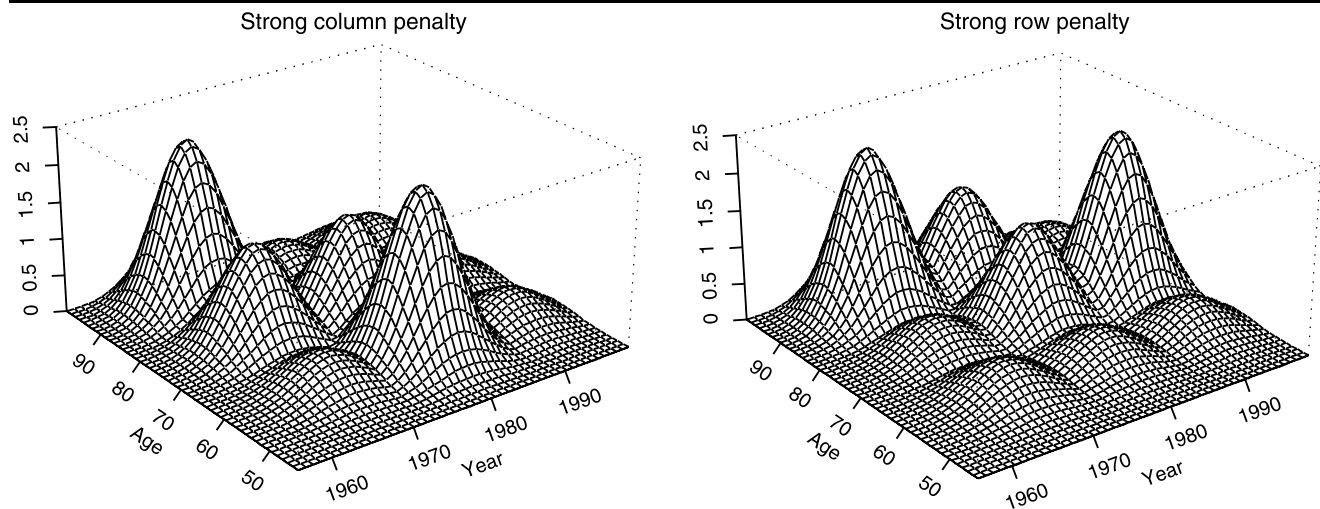
$$G = (B \square B)'W(\check{B} \square \check{B})$$

have a one-to-one correspondence to the elements of  $Q$ . Of course they are arranged differently, because  $Q$  has dimensions  $KL \times KL$  and  $G$  dimensions  $K^2 \times L^2$ . However, it is easy to rearrange the elements of  $G$  to get  $Q$ . Three steps are needed: (1) re-dimension  $G$  to a four-dimensional  $K \times K \times L \times L$  array; (2) permute the second and third dimension; (3) re-dimension to a  $KL \times KL$  matrix.

A similar, but simpler computation finds the right side of (10) by computing and rearranging  $B'(W \cdot Y)\check{B}$ , where  $W \cdot Y$  indicates the element-wise product of  $W$  and  $Y$ .

In a generalized additive model or varying-coefficient model with multiple tensor product bases, weighted inner





**Fig. 8** A visualization of a strong column (*left*) and row (*row*) difference penalty producing linear trends within columns and rows

products of the different bases have to be computed using the same scheme as outlined above. Array regression offers very efficient computation with increases in fitting speed (of far more than 10-fold in most cases) when compared to the following unfolded representation. Typically array regression is used when the data are on a regular grid, however it is possible to include a mix of array and other standard regressors. In such case a hybrid approach is needed, forming Kronecker products like  $(B \square B) \otimes C$  with the design matrix  $C$  of the non-gridded regressors. There will still be efficiency gains, but not as dramatic as for a full grid. An example of a hybrid situation would occur in a model that combines life tables from several countries or regions in one country.

### 3.2 Optimization of the penalty

In the bilinear model, we have four tuning parameters: a penalty on the rows and columns for each of the varying trend and the varying modulation (carrier wave) coefficient surfaces. The penalty in (8) is constructed in such a way that each row (or column) of the tensor product coefficients has the same amount of penalization, but with a breakage in linkage from one row (or column) to the next. Large  $\lambda$  enforces smoothness, whereas small values encourages roughness in either the row or column orientation. Overdispersion was present, thus we chose to optimize the tuning parameters by monitoring and minimizing

$$\text{QIC} = TA + \text{ED} + TA \log(\hat{\phi}),$$

where ED is the approximate effective dimension computed from the trace of the corresponding “hat” or smoother matrix, and  $\hat{\phi} = \text{deviance}(Y, E; \mathcal{A}, \mathcal{B}) / (TA - \text{ED})$ .

We implement the search in a greedy way: each of the four  $\lambda$ s is changed in turn, by one step up and one step down

on the grid. The step that decreases QIC is kept. In the case there is no improvement in QIC, then the current value is kept. In our experience less than 100 sets of  $\lambda$ s are visited, which is an enormous reduction compared to, e.g. with  $K = L = 20$ , a  $20^4$  full grid search.

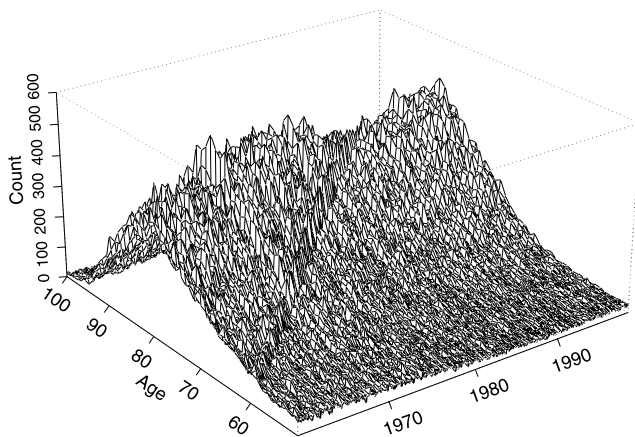
Of course, this is not the most sophisticated way of optimizing the penalty parameters, but is fast and effective. It is also attractive when compared to several alternatives. One could try to exploit the mixed model equivalence (Ngo and Wand 2004), but it is non-trivial to adapt existing mixed model software to use efficient array computations and to handle over-dispersion. The same is true for a fully Bayesian approach, along the lines of Lambert and Eilers (2005).

A modern alternative is boosting (Tutz and Binder 2007), essentially repeating smoothing with too strong a penalty. Array regression can be applied straightforwardly, but one might need to perform many hundreds or even thousands of boosting steps, which compares unfavorably with the efficiency of our grid search.

## 4 Two-dimensional example

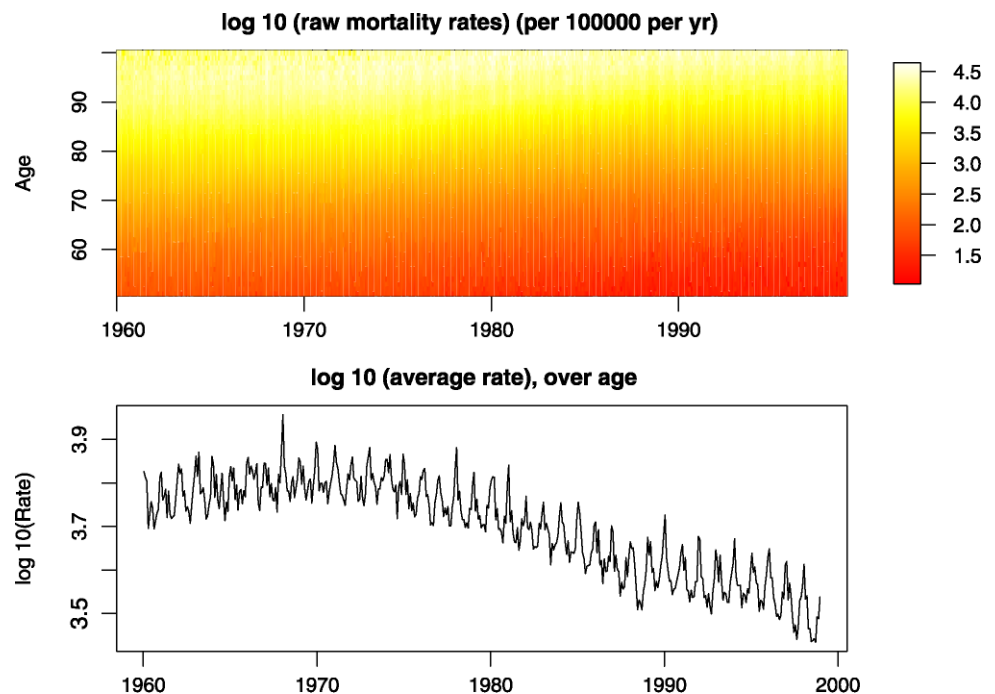
Figure 9 displays the two-dimensional U.S. female monthly raw cerebrovascular death counts, by both year (1960–1998) and age (51–100). This figure can be thought of a visual representation of a large contingency table with 23 400 or  $[50 \times (39 \times 12)]$  cells. A similar perspective plot could be constructed for exposures. As another view, Fig. 10 presents an image plot of cerebrovascular ( $\log_{10}$ ) death rates (per 100 000 per year) (top panel). The bottom panel of this figure shows the ( $\log_{10}$ ) average seasonal trends (averaged over age), which really highlights the strong and varying cyclical behavior from year to year.

We applied the two-dimensional bilinear rate model to these data. Both the trend and the modulation surfaces were constructed on the year and age grid using a basis with  $K \times L = 13 \times 13$  tensor products of (cubic) B-splines; remember that we are smoothing the coefficients and not the raw data. A second order difference penalty was used on both the rows and on the columns of the tensor product coefficients, each having their own tuning parameter to allow anisotropic smoothing. The  $\lambda$ s could take values on a grid with width 0.5 (on a base 10 log scale), and were changed in turn to test for possible downhill turns in (weighted) QIC. Approximately 80 sets of  $\lambda$ s were only visited, opposed to a  $20^4$  full grid search. The results were:  $\lambda = (0.32, 3.2, 10000, 3.2)$ . Figure 11 displays findings for



**Fig. 9** Female cerebrovascular monthly death counts by age (51–100) and year (1960–1998)

**Fig. 10** Image plot of cerebrovascular  $\log_{10}$  death rate (top); average rate over age (bottom)

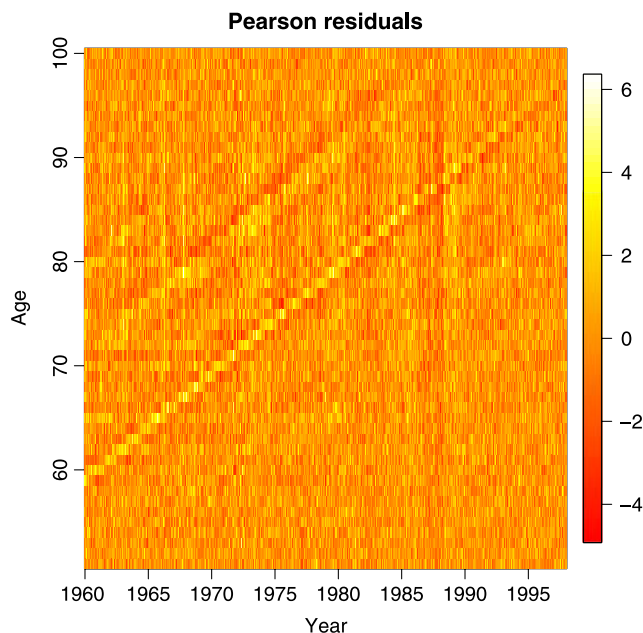
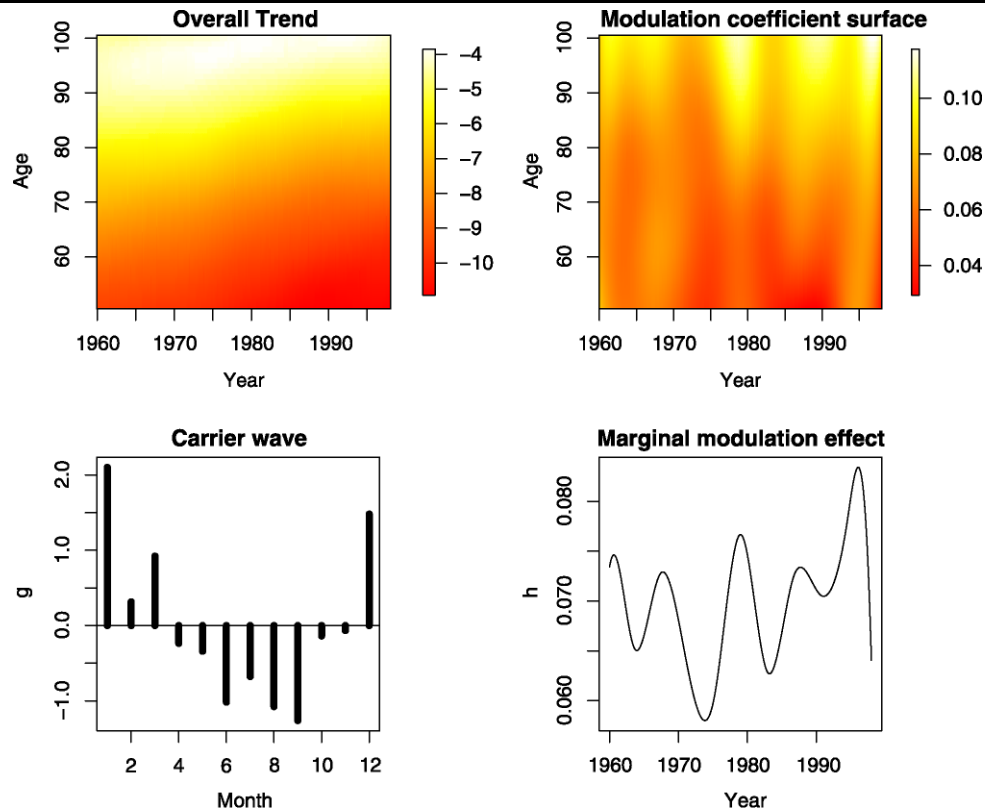


the optimal cerebrovascular model in two-dimensions. The varying-intercept or smooth trend (upper, left) is tracking the raw data as expected. The two-dimensional varying-coefficient surface is displayed for  $H$  (upper, right), as well as its marginalized version across year (lower, right). Note that the large  $\lambda_2 = 3.2 \times 10^4$  suggests a “ribbon” surface in that the modulation effect for fixed time varies in a strongly linear fashion along the age axis. The modulation surface, however, is not additive as the nature of the penalty allows the linear slopes in the direction of age to vary from one time to the next. The carrier wave again shows positive (negative) spikes for winter (summer) months. The combined  $ED = 177.4$ , associated with both estimated varying coefficient surfaces. We see that  $ED < 2 \times 13^2$ , indicating that we chose a sufficiently rich set of bases. Figure 12 displays the Pearson residuals for the two-dimensional bilinear model, which shows many vertical stripes indicating cohort effects. Notice when controlling for age, the (absolute value) size of the Pearson residuals for Fig. 12 is approximately an order of magnitude less when compared to the one-dimensional bilinear model residuals.

## 5 Discussion and extensions

We have proposed bilinear varying-coefficient models for strong seasonal and age dependent effects in series or tables of counts. A beauty of our model is that, by assuming a common carrier-wave over time, we gain parsimony. Yet at the same time, we have not sacrificed generality, as we do not assume a specific functional form for the carrier wave.

**Fig. 11** Female cerebrovascular: Image plot of the varying intercept term or overall trend (*upper, left*); 2D modulation effect for the carrier wave, (*upper, right*); the estimated carrier wave (*lower, left*); Marginal modulation effect (*lower, right*)



**Fig. 12** Cerebrovascular: Image plot of the Pearson residuals for the two-dimensional bilinear model

Beyond understanding and interpreting seasonal effects, certainly there is more future work ahead. The model could perhaps capture even further structure, e.g. serial correlation in the age or time direction, more similar to a more traditional time series model, building on the work of Cai (2007).

Such an approach could prove to be conducive and fruitful toward forecasting beyond the horizon. There also remains the opportunity to model and isolate a “cohort” effect, e.g. as clearly seen along the  $45^\circ$  line in Fig. 12, which poses additional correlation. Further, other approaches to optimizing the tuning parameters could be more fully explored, including: Bayesian through MCMC, mixed models, and boosting.

The bilinear model assumes that the carrier wave has the same pattern for each year and age, only the amplitude changes. We have done some preliminary modelling that relaxes this assumption along the lines of the BAYSEA model (Akaike 1980). Instead of a 12-vector series  $\gamma$ , we introduce a series with as many elements as there are months in the data. To reduce its freedom, penalties are introduced, between identical months in adjacent years. This forces the parameter value for, e.g., May 1990 not to deviate too much from the average of the parameter values for May 1989 and May 1991. This scheme is applied to all years and all months. In one dimension the model is  $\eta_t = v_t + s_t$ , where  $v$  is the trend,  $s$  is the seasonal component. A second order difference penalty is put on  $v$ . The penalty on  $s$  is seasonal, having components  $(s_1 - 2s_{13} + s_{25})^2 + (s_2 - 2s_{14} + s_{26})^2 + \dots$ . Thus the elements of  $s$  corresponding to, e.g., January readings get a second order penalty, as do other months. Essentially this model contains one long series, the trend  $v$ , and 12 shorter smooth series, one for each month. To gain effi-

ciency in computation, we propose to use P-spline smoothing for  $v$  as well as for each of the twelve smooth series corresponding to the months. This model is more flexible than the periodic carrier wave modulation model, but is more difficult to interpret. In one dimension, the BAYSEA model has no modulation component. In two dimensions we can imagine a carrier wave in BAYSEA style, which is modulated by a surface along age and time.

It is interesting to compare our model to some other (one-dimensional) approaches. Harvey (1989) discussed cyclical and seasonal state space models. Fahrmeir and Tutz (2001) further adapted this work to a generalized linear model setting. To simplify the discussion we outline the linear case. The model represents a time series by  $y_i = t_i + \gamma_i + e_i$ , where  $t$  represents trend,  $\gamma$  the seasonal component and  $e$  noise. The objective function is

$$S = \|y - t - \gamma\|^2 + \lambda \|Dt\|^2 + \kappa \|P\gamma\|^2.$$

Here  $D$  is the usual matrix for forming differences of order  $d$ . The matrix has a special structure to push the sum of 12 (if the period is 12) adjacent values of  $\gamma$  towards zero. In row  $i$  of  $P$ , the elements from  $i$  to  $i + 11$  are 1, while all other elements in that row are zero. This interesting penalty gives  $\gamma$  freedom to fit a seasonal deviation from the trend; the larger the  $\kappa$ , the more similar the elements of  $\gamma$ —that are 12 observations apart—will be. Although Akaike's BAYSEA model does something similar to these other approaches, it penalizes (higher order) differences between elements of  $\gamma$  that are 12 months apart. In this way, the penalty can allow for a more gradual change in the seasonal patterns over time.

The purpose of our bilinear model is to estimate one periodic pattern that only changes in strength, but not in shape, over time. This evidently simplifies interpretation over Akaike's BAYSEA or state space model because one only has to look at twelve numbers when studying the seasonal pattern. Naturally, whether the bilinear model works

well or not depends on the data; we found it to be successful for the mortality data that we studied.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

## References

- Akaike, H.: Seasonal adjustment by a Bayesian modeling. *J. Time Ser. Anal.* **1**, 1–14 (1980)
- Cai, Z.: Trend time-varying coefficient time series models with serially correlated errors. *J. Econom.* **136**, 163–188 (2007)
- Currie, I.D., Durbán, M., Eilers, P.H.C.: Generalized linear array models with applications to multidimensional smoothing. *J. R. Stat. Soc., Ser. B* **68**, 259–280 (2006)
- Eilers, P.H.C., Currie, I.D., Durbán, M.: Fast and compact smoothing on large multidimensional grids. *Comput. Stat. Data Anal.* **50**, 61–76 (2006)
- Eilers, P.H.C., Gampe, J., Marx, B.D., Rau, R.: Modulation models for seasonal time series and incidence tables. *Stat. Med.* **27**(17), 3430–3441 (2008)
- Eilers, P.H.C., Marx, B.D.: Generalized linear additive smooth structures. *J. Comput. Graph. Stat.* **11**(4), 758–783 (2002)
- Fahrmeir, L., Tutz, G.: *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd edn. Springer, New York (2001)
- Harvey, A.C.: *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge (1989)
- Lambert, P., Eilers, P.H.C.: Bayesian proportional hazards model with time varying regression coefficients: A penalized Poisson regression approach. *Stat. Med.* **24**, 3977–3989 (2005)
- Lee, Y., Nelder, J.A., Pawitan, Y.: *Generalized Linear Models with Random Effects*. Chapman and Hall/CRC Press, London (2006)
- Ngo, L., Wand, M.P.: Smoothing with mixed model software. *J. Stat. Softw.* **9**(1), 1–54 (2004)
- Rau, R., Doblhammer, G.: Seasonal mortality in Denmark. The role of sex and age. *Demogr. Res.* **9**, 197–222 (2003)
- Tutz, G., Binder, H.: Flexible modelling of discrete failure time including time-varying smooth effects. *Stat. Med.* **23**(15), 2445–2461 (2004)
- Tutz, G., Binder, H.: Boosting ridge regression. *Comput. Stat. Data Anal.* **51**(12), 6044–6059 (2007)